Machine Learning 6.867 - Project

December 10, 2015

1 Introduction

Logistic regression is a widely-used method for classification. Given training data $(\mathbf{x}_i, y_i)_{i=1}^n$ with features $\mathbf{x}_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, +1\}$, we would like to find $P(y = 1|\mathbf{x})$ for an arbitrary vector $\mathbf{x}$. For logistic regression, we assume a logistic model, where $P(y = 1|\mathbf{x}) = \dfrac{1}{1 + e^{-y\mathbf{w}^T\mathbf{x}}}$, and $\mathbf{w} \in \mathbb{R}^p$ is a weight vector. To determine the value of the weight vector, we can do maximum-likelihood estimation, and optimize to find the value of $\mathbf{w}$ which maximizes the negative log-likelihood. This results in a concave, nonlinear optimization problem solvable via gradient descent.

Alternatively, we might take the Bayesian perspective and assume a prior on $\mathbf{w}$. This is preferable for cases where we interested in the covariance of the weight vector and/or the predicted probabilities of new data points. However, this is a difficult problem and there is no simple analytic formula to compute the posterior and predictive distributions directly. Therefore, in this case we turn to variational methods in order to obtain a close approximation to the posterior for $\mathbf{w}$. From this, we can then obtain estimates for the predictive distribution $P(y = 1|\mathbf{x})$.

In this paper, we develop and test an implementation of Mean-Field Variational Bayes logistic regression based on previous literature. Paper [2] provides a walkthrough of MFVB applied to logistic regression, and sample MATLAB code is available on Github. We develop our own implementation based off of this package of code, and we code these functions in Julia. Going beyond this work, we compare our method against Markov-Chain Monte-Carlo, and implemented this method using an existing package in `R`. We performed sensitivity analysis varying the hyperpriors of $\mathbf{w}$ to generate the data, and we obtain a variety of different simulated datasets. In the section of computational results, we present train and test set accuracy for all of the simulated datasets, for standard logistic regression, MFVB, and MCMC. We obtain estimates for the predicted probabilies of MCMC by taking the average of the weight vector $\mathbf{w}$ over all iterations, and then use that to compute the Bernoulli probabilities. In addition, we also present detailed MCMC results for Dataset 0, along with comparisons to MFVB logistic regression.

2 Mean-Field Variational Bayes

Mean-field variational Bayes (MFVB) is a method for approximating the posterior distribution. In general, we have unknown parameters $w_1, w_2, \ldots, w_n$ that we have priors on, and our objective is to find the joint distribution $p(w_1, w_2, \ldots, w_n)$. Assuming that our approximate distribution is in the family $Q = \{q : q(w_1, w_2, \ldots, w_n) = q(w_1)q(w_2)\ldots q(w_n)\}$, we find $q^* \in Q$ that minimizes the KL-divergence with p, i.e. $q^* = \min KL(q||p)$. In particular, for logistic regression, the analytical form of the posterior is unknown and has been approximated with MFVB in the literature [2]. We use local variational bounds on the conditional probability using the convexity of the logarithm function. In particular, we use a variational treatment based on the approach of Jaakkola and Jordan (2000). This approach consists of approximation the likelihood function of the logistic regression, governed by the sigmoid function, by the exponential of the a quadratic form, leading to a gaussian approximation of the posterior distribution. More explicitly, if $y \in \{-1, 1\}$ is a target variable for a data vector

x then the likelihood function of the target variable y is:

$$p(y|x, w) = \sigma(yw^T x) \tag{1}$$

with w being the logistic regression weight, and $\sigma(x) = \dfrac{1}{1+\exp(-x)}$ the sigmoid function. Using a transformation of a the logarithm of the sigmoid and the concept of convex duality, we get:

$$\sigma(x) \geq \sigma(\xi)\exp((x-\xi)/2 - \lambda(\xi)(x^2 - \xi^2)) \tag{2}$$

where

$$\lambda(\xi) = \frac{1}{2\xi}[\sigma(\xi) - \frac{1}{2}] = \frac{1}{4\xi}\tanh(\frac{\xi}{2})$$

and ξ is a variational parameter.

Therefore, if we let $a = w^T x$ we get:

$$p(y|x, w) \geq e^{ya}\sigma(\xi)\exp\{-(\xi + a)/2 - \lambda(\xi)(a^2 - \xi^2)\} \tag{3}$$

To every training set observation (x_n, y_n), there is a variational parameter ξ_n associated. We apply the bound above to each of the terms in the likelihood function. Let $Y = [y_1, y_2, \ldots, y_n]^T$ and the X be the data matrix, then the likelihood function is:

$$p(Y|X, w) = \prod_{i=1}^{N} p(y_i|x_i, w) = \prod_{i=1}^{N} \sigma(yw^T x) \tag{4}$$

and thus we obtain the following bound on the joint distribution on y and w, assuming a prior $p(w)$ on w:

$$p(Y, w|X) = p(Y|X, w)p(w) \geq h(w, \xi)p(w) \tag{5}$$

and $h(w, \xi) = \prod_{i=1}^{N} e^{y_i w^T x_i}\sigma(\xi_i)\exp\{-(\xi_i + w^T x_i)/2 - \lambda(\xi_i)((w^T x_i)^2 - {\xi_i}^2)\}$

However, the variational Bayes approximation is known to underestimate the covariance matrix of the posterior distribution, and this estimate can be made arbitrarily bad for simulated examples with 2 or more dimensions [1].

3 Implementation

3.1 MFVB Logistic Regression

3.2 Simulated Datasets

4 Markov-Chain Monte Carlo

To evaluate the quality of the covariance estimates produced by our method, we used Markov-Chain Monte Carlo (MCMC) as a benchmark for the "true" distribution of the logistic regression weights (w_0, w_1, w_2). We used the `R` package `MCMCpack` with an improper uniform prior, 10,000 iterations, and burn-in rate of 1,000 iterations. We also compared to an MCMC simulation with a normal prior on the weights $\mathcal{N}(\mathbf{0}, 1000I)$ and found similar results. Figures 1 and 2 show the progression of the MCMC algorithm assuming each prior.

To visualize the joint distribution of the logistic regression weights, we plot the MCMC results for the values of w_1 and w_2. In addition, we fit the kernel density to the

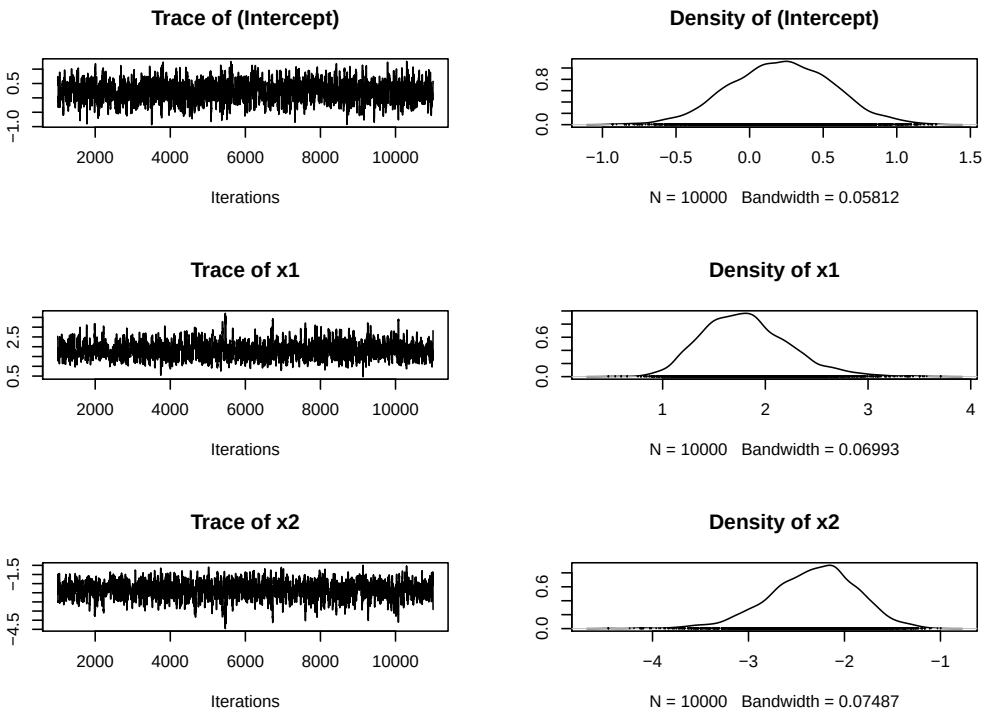

Figure 1: MCMC simulations of logistic regression weights for dataset 1, and corresponding marginal density plots, assuming an improper uniform prior. 10,000 iterations total.

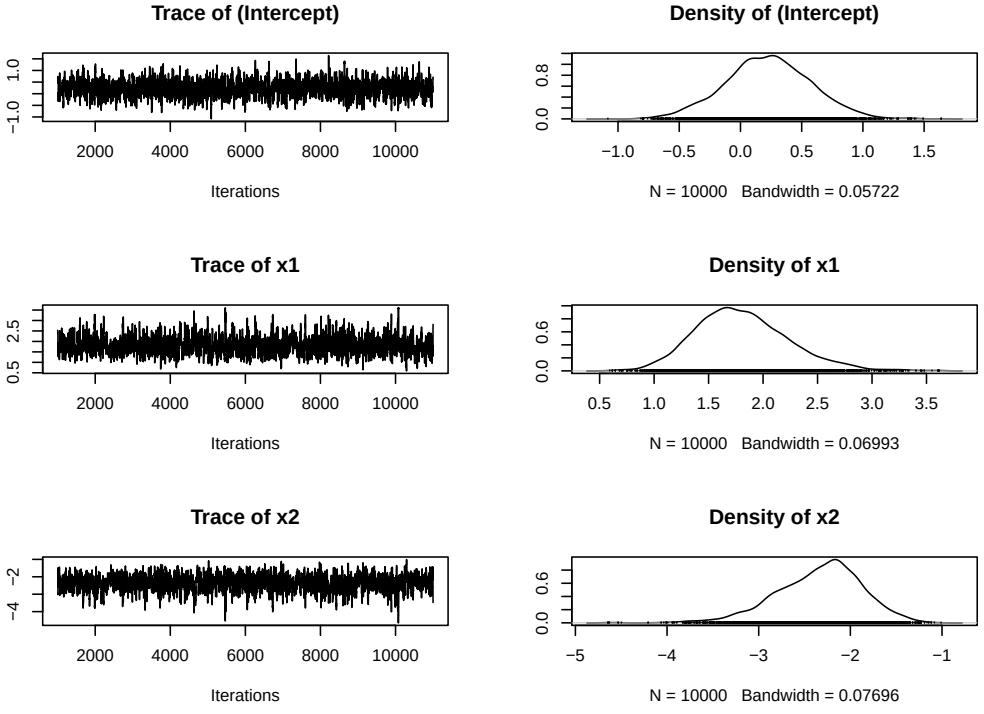

Figure 2: MCMC simulations of logistic regression weights for dataset 1, and corresponding marginal density plots, assuming a normal prior $\mathcal{N}(\mathbf{0}, 1000I)$. 10,000 iterations total.

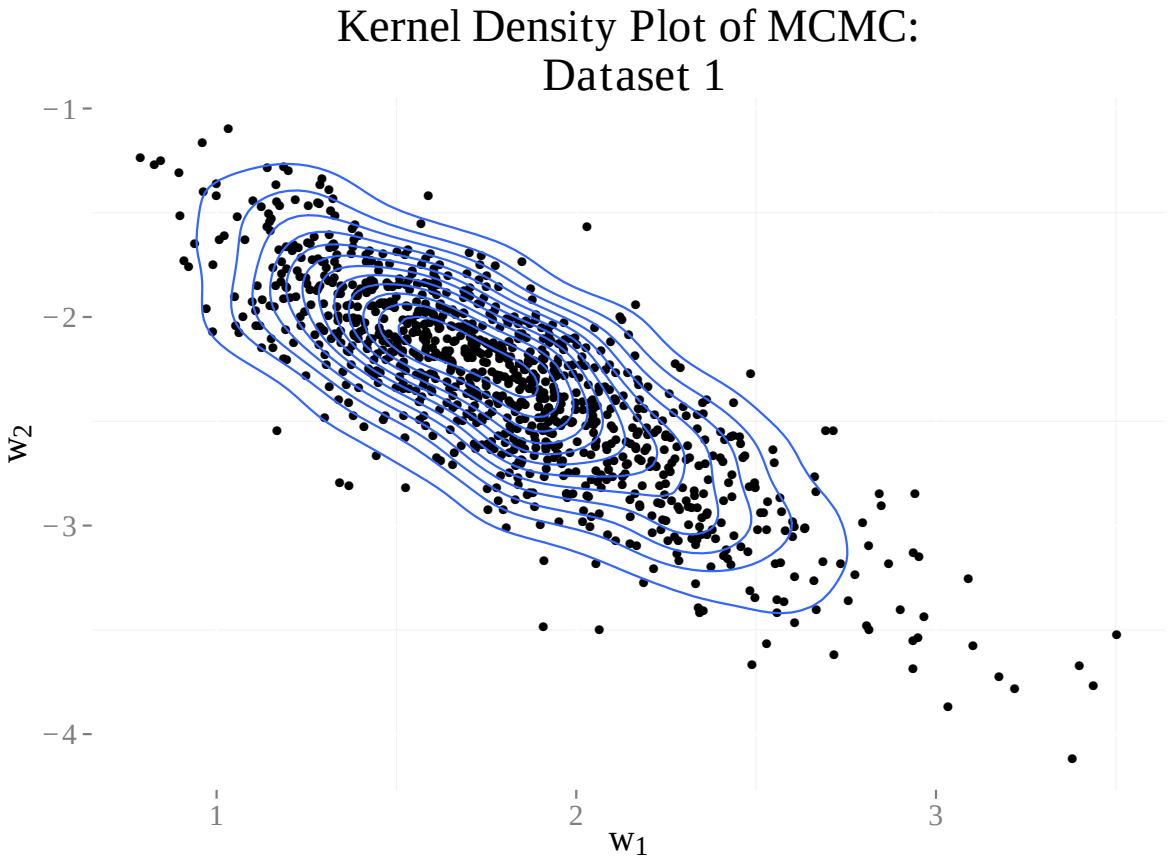

Figure 3: MCMC simulations of logistic regression weights for dataset 1, and corresponding kernel density plot, assuming an improper uniform prior. Subset of 1,000 out of 10,000 total iterations shown.

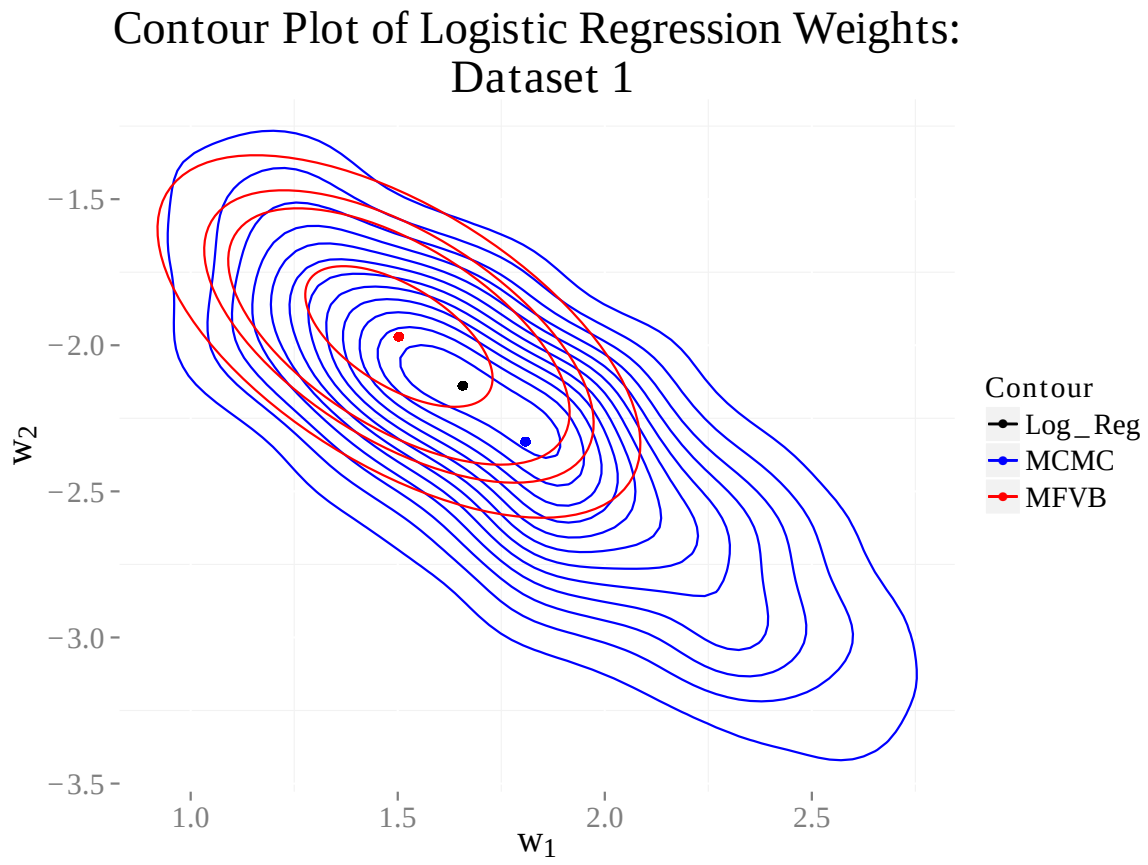

Figure 4: Comparison of logistic regression point estimate, kernel density plot of MCMC simulations, and posterior density of MFVB logistic regression (vectorized function), for dataset 1. Contours of MFVB logistic regression indicate 50%, 90%, 95%, and 99% confidence intervals for the bivariate normal $\mathcal{N}(\mathbf{w}_N, \mathbf{V}_N)$. Mean values for MCMC and MFVB are also included.

Dataset	Logit Train	Logit Test	MCMC Train	MCMC Test	MFVB Train	MFVB Test
0	0.8700	**0.8580**	0.8700	**0.8580**	0.8600	0.8570
1	0.7500	0.8450	0.7500	**0.8460**	0.7400	0.8450
2	0.5500	**0.4890**	0.5100	**0.4890**	0.5400	0.4750
3	0.6900	**0.7420**	0.6900	**0.7420**	0.7200	0.7120
4	0.8600	**0.8050**	0.8600	**0.8050**	0.8600	0.8040
5	0.9400	0.9290	0.9400	**0.9300**	0.9400	0.9260
6	0.7400	0.7130	0.7400	**0.7140**	0.6900	0.6840
7	1.0000	0.9760	1.0000	***	1.0000	**0.9780**
8	0.9200	**0.8670**	0.9200	**0.8670**	0.9300	0.8640
9	0.5200	0.3980	0.5100	0.4020	0.4600	**0.4130**
10	0.6700	0.6690	0.6700	**0.6700**	0.6300	0.6560

Table 1: Training and test set accuracy for logistic regression, MCMC logistic regression, and MFVB logistic regression on all datasets. The highest out-of-sample accuracy score is highlighted for each dataset.
*** : The training dataset was completely separable; therefore the posterior covariance matrix was zero and the precision matrix was undefined. Thus, we did not obtain MCMC predictive probabilities for this dataset.

5 Future Work

A recent paper by Giordano, Broderick and Jordan (2015) proposes a linear correction term derived analytically to improve the covariance estimate for MFVB. The new proposed method is called Linear Response Variational Bayes (LRVB). In the beginning of this project, we considered applying this new method on logistic regression and comparing its performance to the benchmark of MFVB and regular logistic regression. However, we ran out of time to complete this phase, because implementing MFVB for logistic regression was challenging and generating the plots for the MCMC vs MFVB comparison was difficult. In the future, we are planning to experiment with LRVB applied to variational Bayesian logistic regression, because we have majority of the most difficult work completed and now we have the source code for this method. From this research, we hope to develop an improved version of MFVB logistic regression.